KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Business and Economics

A comparison of robust versions of the AIC based on M, S and MM-estimators

Kukatharmini Tharmaratnam and Gerda Claeskens

DEPARTMENT OF DECISION SCIENCES AND INFORMATION MANAGEMENT (KBI)

KBI 1014

A comparison of robust versions of the AIC based on M, S and MM-estimators

Kukatharmini Tharmaratnam and Gerda Claeskens

OR & Business Statistics and Leuven Statistics Research Center

K.U.Leuven, Belgium

June 25, 2010

Abstract

Variable selection in the presence of outliers may be performed by using a robust version of Akaike's information criterion AIC. In this paper explicit expressions are obtained for such criteria when S and MM-estimators are used. The performance of these criteria is compared to the existing AIC based on M-estimators and to the classical non-robust AIC. In a simulation study and in data examples we observe that the proposed AIC with S and MM-estimators selects more appropriate models in case outliers are present.

Keywords: Akaike's information criterion; robust estimation; variable selection.

1 Introduction

It has been recognized that variable selection procedures need special care in the presence of outliers in the data. Since most of the classical procedures are likelihood-based, alternatives have been developed. Some of the main developments to make classical model selection procedures for linear models less sensitive to outlying observations are a robust version of Akaike's information criterion (AIC Akaike, 1973) based on M estimators (Ronchetti, 1985), a robust C_p (Ronchetti and Staudte, 1994; Sommer and Staudte, 1995), a robust version of cross-validation (Ronchetti et al., 1997), see also the survey presented in Ronchetti (1997). Qian and Künsch (1998) select models in a robust way using the concept of stochastic complexity, while Agostinelli (2002) rather deals with weighted versions of likelihood estimators. Several of these model selection methods are described in Maronna et al. (2006, Sec. 5.12) and Claeskens and Hjort (2008, Ch. 2 and 4). Müller and Welsh (2005) make use of the bootstrap to combine a robust penalized criterion with a robust conditional expected prediction loss function. Other use of the bootstrap

for robust variable selection is made by Salibián-Barrera and Van Aelst (2008). Heritier et al. (2009, p. 159) present a form of the AIC based on robust quasilikelihood.

While the emphasis in the existing literature is mostly on M-estimation when it comes to variable selection methods, in this paper we investigate whether improvements can be achieved when using S or MM-estimators. The derivation of information criteria in the style of the AIC using these robust estimators is in the line with the generalized information criteria of Konishi and Kitagawa (1996). When applied to estimation in likelihood models free of outliers, this approach would lead to Takeuchi's information criterion (Takeuchi, 1976), which differs from the traditional AIC only in its penalty term.

The rest of the paper is organized as follows. Section 2 gives the definition of the robust estimators that are used within the information criteria, in particular, we use M-estimators, (uniform) S and MM-estimators (Omelka and Salibián-Barrera, 2008) for linear regression models. Section 3 introduces and derives the formulae for the robust version of AIC based on each of these different estimators. Section 4 reports the results of a simulation study and data examples that compares the performance of classical AIC, AIC based on M, S and MM-estimation. Finally, section 5 contains a discussion and concluding remarks. The appendix contains the `R` software that is used for the calculations.

2 Robust estimation methods

We consider the linear regression model

$$Y_i = \theta_0^t X_i + u_i, \quad i = 1, \ldots, n, \tag{1}$$

where the response variables $Y_i \in \mathbb{R}$ $(i = 1, \ldots, n)$ are independent, the covariate vector $X_i \in \mathbb{R}^p$ with a corresponding coefficient vector $\theta_0 \in \mathbb{R}^p$ and the u_i are random errors independent from the explanatory variable X_i, with mean zero and constant variance σ^2. For normal errors with standard deviation σ, the Akaike information criterion for variable selection is given by

$$\text{AIC} = 2n \log \hat{\sigma} + 2(p+1) + \{n + n \log(2\pi)\}, \tag{2}$$

where the last term, $\{n + n \log(2\pi)\}$, may be omitted because it is independent of the choice of variables in the model and where $\hat{\sigma}$ is the maximum likelihood estimate of σ. The AIC is defined

as minus twice the value of the maximized log likelihood plus twice the number of estimated parameters in the model. The penalty takes the p regression coefficients θ_0 and the unknown error variance into account.

In the case that outliers are present in the data, only the majority of the data follows the above model (1). Extreme observations might occur in both the explanatory variables and the response. It is in these circumstances that we wish to investigate the inclusion or exclusion of components of the covariate vector X.

We first give an overview of some robust estimation methods before turning to versions of the AIC that are based on these estimation methods.

2.1 M-estimators

A general M-estimator Huber (2004) is defined as the minimum with respect to θ of the objective function $\sum_{i=1}^{n} \rho(y_i|x_i, \theta)$, for a given function ρ that has the properties of being even, non-decreasing in $[0, \infty)$ and with $\rho(0) = 0$. Equivalently, when the response values $Y_1, \ldots, Y_n$ are independent, the M-estimator for θ solves the equation

$$\sum_{i=1}^{n} \psi(y_i|x_i, \theta) = 0 \tag{3}$$

where $\psi(y|x, \theta) = \frac{\partial}{\partial\theta}\rho(y|x, \theta)$. Intuitively, to take care of outliers which result in large residuals, the function $\rho(\cdot)$ should increase at a slower rate than t^2, particularly for large residuals. A common choice for ρ is given by Huber's family with an unbounded loss function

$$\rho_c(t) = \begin{cases} t^2 & \text{if } |t| \leq c \\ 2\,c\,|t| - c^2 & \text{if } |t| > c\,, \end{cases} \tag{4}$$

where $c > 0$ is a tuning constant that can be thought of as a threshold value such that observations with residuals larger than c have a reduced effect in the estimating equation (3). A value of 95% asymptotic efficiency on the standard normal distribution is obtained when the constant equals 1.345 (Huber, 2004). In practice, a typical choice for c is 1.345 $\widehat{\sigma}_m$, with $\widehat{\sigma}_m$ the median absolute deviation (MAD) of the residuals, $\text{MAD}(r_1, \ldots, r_n) = 1.4826\ \text{median}_{i=1,\ldots,n}(|r_i - \text{median}(r_1, \ldots, r_n)|)$. The M estimator is computed with $\rho(y_i|x_i, \theta) = \rho_c\left(\frac{y_i - \theta^t x_i}{\widehat{\sigma}_m}\right)$. In practice, iteration is used between estimation of θ and of σ until convergence.

2.2 S-estimators

S-estimators for linear regression were introduced by Rousseeuw and Yohai (1984) as an alternative to M estimators that do not suffer that much from leverage points (which are outliers in the covariates) and at the same time have a high breakdown point and do not require an auxiliary scale estimator.

Let G_0 and F_0 be the cumulative distribution functions of X_i and u_i respectively. The cumulative distribution of (Y_i, X_i) under model (1) is then given by $H_0(y, x) = G_0(x)F_0(y-\theta_0^t x)$. In the presence of outliers, we make the assumption that the cumulative distribution function H of the data belongs to a contamination neighborhood of H_0 of size ϵ_0. More precisely,

$$H \in \mathcal{H}_{\epsilon_0} = \{(1-\epsilon)H_0 + \epsilon H^*; \epsilon \in [0, \epsilon_0]\},$$

where H^* is an arbitrary cumulative distribution function and $\epsilon_0 < 0.5$.

The loss function ρ_0 is a function that is even, continuously differentiable, non-decreasing on $[0, \infty)$, satisfies that $\rho_0(0) = 0$ and has $\sup_{u\in\mathbb{R}} \rho_0(u) = 1$. We define $b = E_{F_0}[\rho_0(u_1)]$, with u_1 one of the error terms in model (1), and assume that $\epsilon_0 < b < 1 - \epsilon_0$ to ensure consistency of the scale estimator under the central model F_0. The notation E_{F_0} means that the expectation is computed with respect to F_0.

First we implicitly define the scale function $\widehat{\sigma}_n(\theta)$ by that function of θ that satisfies the equation

$$\frac{1}{n}\sum_{i=1}^{n} \rho_0\left(\frac{y_i - \theta^t x_i}{\widehat{\sigma}_n(\theta)}\right) = b, \tag{5}$$

and take $\rho(y_i|x_i, \theta) = \rho_0\left(\frac{y_i-\theta^t x_i}{\widehat{\sigma}_n(\theta)}\right)$. The S-estimator $\widehat{\theta}_s$ minimizes the scale function, that is, $\widehat{\theta}_s = \text{argmin}_{\theta\in\mathbb{R}^p}\widehat{\sigma}_n(\theta)$, and the scale estimator itself is $\widehat{\sigma}_s = \widehat{\sigma}_n(\widehat{\theta}_s)$.

Omelka and Salibián-Barrera (2008) obtain the uniform consistency of the S-estimator over the contamination neighborhood $\mathcal{H}_{\epsilon_0}$ and assume thereby that F_0 is absolutely continuous with an even and positive density function over the real line. More precisely, using notation that we will return to when constructing a version of the AIC for variable selection, for each $\theta \in \mathbb{R}^p$ and $H \in \mathcal{H}_{\epsilon_0}$, define a scale function $\sigma(H, \theta)$ that satisfies

$$E_H\big[\rho_0\left(\frac{Y_1 - \theta^t X_1}{\sigma(H, \theta)}\right)\big] = b,$$

where E_H is the expectation computed with respect to H. Hence, for each $\theta \in \mathbb{R}^p$ we define a functional $\sigma(., \theta) : \mathcal{F} \to \mathbb{R}_+$, with domain $\mathcal{F} \supset \mathcal{H}_{\epsilon_0}$. The associated functional S estimators of

location and scale are then defined as

$$\theta_s(H) = \arg \inf_{\theta \in \mathbb{R}^p} \sigma(H, \theta), \quad \sigma_s(H) = \inf_{\theta \in \mathbb{R}^p} \sigma(H, \theta).$$

Under certain regularity conditions (see Omelka and Salibián-Barrera, 2008) the S-estimators of the regression parameters ($\widehat{\theta}_s$) and scale ($\widehat{\sigma}_s$) are consistent estimators of the functionals $\theta_s(H)$ and $\sigma_s(H)$ respectively. Theorems 1 and 2 in Omelka and Salibián-Barrera (2008) show that $\widehat{\theta}_s$ and scale $\widehat{\sigma}_s$ are uniformly consistent over the whole contamination neighbourhood $\mathcal{H}_{\epsilon_0}$.

A commonly used family of loss functions ρ_0 is given by Tukey's bi-square family (Beaton and Tukey, 1974)

$$\rho(u; d) = \begin{cases} 3\,(u/d)^2 - 3\,(u/d)^4 + (u/d)^6 & \text{if } |u| \leq d\,, \\ 1 & \text{if } |u| > d\,. \end{cases} \tag{6}$$

The choice $d = 1.5476$ yields $b = E_\Phi\left[\rho\left(Z; d\right)\right] = 0.5$, with Φ the standard normal cumulative distribution function and $Z \sim N(0, 1)$. The associated S-regression estimator has maximal asymptotic breakdown point 50% (Rousseeuw and Yohai, 1984). Estimators with 30% breakdown point are gotten when $d = 2.5608$, resulting in a higher efficiency. Both options are contrasted in the simulation study.

2.3 MM-estimators

A further step in robust estimation uses the S-scale estimator in an M-estimation equation. Let $\rho_1 : \mathbb{R} \to \mathbb{R}_+$ be another loss function such that $\rho_1(u) \leq \rho_0(u)$ for all $u \in \mathbb{R}$ and $\sup_u \rho_1(u) = \sup_u \rho_0(u)$. The MM-regression estimator $\widehat{\theta}_{mm}$ is defined the global minimum of $f : \mathbb{R}^p \to \mathbb{R}_+$, with

$$f(\theta) = \frac{1}{n} \sum_{i=1}^{n} \rho_1 \left(\frac{y_i - \theta^t x_i}{\widehat{\sigma}_s} \right).$$

Thus,

$$\widehat{\theta}_{mm} = \text{argmin}_{\|\theta\| \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho_1 \left(\frac{y_i - \theta^t x_i}{\widehat{\sigma}_s} \right),$$

with the corresponding functional

$$\theta_{mm}(H) = \text{argmin}_{\|\theta\| \in \mathbb{R}^p} E_H \left[\rho_1 \left(\frac{Y_i - \theta^t X_i}{\sigma_s(H)} \right) \right].$$

Omelka and Salibián-Barrera (2008) prove the consistency and the uniform asymptotic normality of such MM-estimators. The MM-variance estimator is taken to be the S scale estimator $\widehat{\sigma}_s$. In practice $\rho_1 = \tilde{\rho}_d$ is often a re-scaled version of $\rho_0 = \rho_d$ (Tukey's bi-square family loss function).

We will use all of these definitions in the following sections.

3 Robust versions of the AIC

Akaike's information criterion is in full likelihood models defined as AIC $= -2$ log-likelihood$(\widehat{\theta})+$ 2 length(θ), with length(θ) the number of parameters that are estimated in the model, and with $\widehat{\theta}$ the maximum likelihood estimator of the model parameters θ. The AIC arises as an estimator of the expected value of the Kullback-Leibler distance between the maximized density of the data implied by the model and the true data generating density g, that is nearly always unknown,

$$\mathrm{KL}(g, f(.,\widehat{\theta})) = \int\int g(y|x)\log g(y|x)dydG(x) - R_n$$

where $R_n = \int\int g(y|x)\log f(y|x,\widehat{\theta})dydG(x)$ and G is the cumulative distribution function of X. A derivation of the traditional AIC can for example be found in Claeskens and Hjort (2008, Sec. 2.3). Since the AIC is likelihood-based, and thus is sensitive to outlying observations in the data, we here search for more robust alternatives, in the spirit of the generalized information criterion of Konishi and Kitagawa (1996).

3.1 Derivation of a robust AIC

Instead of working with the maximized likelihood function in the Kullback-Leibler distance, we use the loss function ρ and the corresponding robust estimator $\widehat{\theta}$ and consider as a good model one that minimizes the expected value of the following weighted Kullback-Leibler distance that involves the empirical distribution of the covariates,

$$\frac{1}{n}\sum_{i=1}^{n}\int g(y|x_i)\{\log g(y|x_i) + \rho(y|x_i,\widehat{\theta})\}dy = \frac{1}{n}\sum_{i=1}^{n}\int g(y|x_i)\log g(y|x_i)dy + R_n^{\rho}, \qquad (7)$$

where $R_n^{\rho} = \frac{1}{n}\sum_{i=1}^{n}\int g(y|x_i)\rho(y|x_i,\widehat{\theta})dy$. In the next section we make this more concrete for the different robust estimators. For M-estimators, such a robust AIC has been obtained by **?**.

Since the first term is independent of the model, the key quantity to study is R_n^{ρ}, which depends on the data through the robust estimator $\widehat{\theta}$. The expected value of R_n^{ρ} with respect to the robust estimator, under the true density g for the response variable Y_i given the covariate is equal to

$$Q_n = E(R_n^{\rho}) = \frac{1}{n}\sum_{i=1}^{n} E\left[\int g(y|x_i)\rho(y|x_i,\widehat{\theta})dy\right],$$

which is estimated by replacing the true distribution functions by their empirical counterparts, leading to the estimator

$$\widehat{Q}_n = \frac{1}{n}\sum_{i=1}^{n} \rho(Y_i|x_i, \widehat{\theta}).$$

For maximum likelihood estimation, $\widehat{Q}_n$ corresponds to the minus log likelihood function, evaluated at the maximum likelihood estimator, divided by the sample size. To construct an AIC, we investigate the bias of $\widehat{Q}_n$ for estimation of Q_n, which will lead to an appropriate penalty term in the variable selection criterion.

Define by $\theta_{0,n}$ the least false parameter vector that corresponds to the empirical distribution of the covariates and thus maximizes $\mathcal{Q}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \int g(y|x_i)\rho(y|x_i,\theta)dy$. Denote $Q_{0,n} = \mathcal{Q}_n(\theta_{0,n})$, $V_n = \sqrt{n}(\widehat{\theta} - \theta_{0,n})$ and $J_n = -\frac{1}{n}\sum_{i=1}^{n} \int g(y|x_i) I(y|x_i, \theta_{0,n})dy$, with information function $I(y|x,\theta) = -\frac{\partial^2 \rho(y|x,\theta)}{\partial\theta\partial\theta^t}$. The score function is defined as $u(y|x,\theta) = -\frac{\partial \rho(y|x,\theta)}{\partial\theta}$, with variance $K_n = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}\{u(Y|x_i,\theta_{0,n})\}$. The limit versions of J_n and K_n are denoted by J and K, respectively.

Proposition 1 *Let $\bar{Z}_n$ be the average of the values $Z_i = -\rho(Y_i|x_i,\theta_{0,n}) + \int g(y|x_i)\rho(y|x_i,\theta_{0,n})dy$, assume that ρ is two times differentiable, and using the notation as defined above,*

$$\widehat{Q}_n - R_n^{\rho} = -\bar{Z}_n - \frac{1}{n}V_n^t J V_n + o_p(1/n). \tag{8}$$

Proof. A Taylor expansion for R_n^{ρ} gives that

$$\begin{aligned} R_n^{\rho} &= \frac{1}{n}\sum_{i=1}^{n} \int \Big\{ g(y|x_i) \Big[\rho(y|x_i,\theta_{0,n}) - u(y|x_i,\theta_{0,n})(\hat{\theta} - \theta_{0,n}) \\ &\quad -\frac{1}{2}(\hat{\theta} - \theta_{0,n})^t I(y|x_i,\theta_{0,n})(\hat{\theta} - \theta_{0,n}) + o_P(1/n) \Big] \Big\} dy = Q_{0,n} + \frac{1}{2n}V_n^t J_n V_n + o_P(1/n). \end{aligned}$$

In a similar fashion, a Taylor expansion for $\widehat{Q}_n$ results in

$$\begin{aligned} \widehat{Q}_n &= \frac{1}{n}\sum_{i=1}^{n} \left\{ \rho(Y_i|x_i,\theta_{0,n}) - u(Y_i|x_i,\theta_{0,n})(\hat{\theta} - \theta_{0,n}) - \frac{1}{2}(\hat{\theta} - \theta_{0,n})^t I(Y_i|x_i,\theta_{0,n})(\hat{\theta} - \theta_{0,n}) \right\} \\ &\quad + o_P(1/n) = Q_{0,n} - \bar{Z}_n - \frac{1}{2n}V_n^t J_n V_n + o_P(1/n). \end{aligned}$$

Thus, it holds that $\widehat{Q}_n - R_n^{\rho} = -\bar{Z}_n - \frac{1}{n}V_n^t J_n V_n + o_P(1/n)$. □

From (8) and since for robust estimators it holds that $V_n \xrightarrow{d} N(0, J^{-1}KJ^{-1})$, it follows that $E(\widehat{Q}_n - Q_n)$ is approximately (leaving out remainder terms of smaller order) equal to

$-\text{Trace}(J^{-1}K)/n$. Based on these results, a model selection criterion in the style of Akaike's information criterion is to compute $\widehat{Q}_n + \text{Trace}(J_n^{-1}K_n)/n$ for each candidate model, and then to select the model with the smallest such value. Equivalently, we define a robust AIC, specific to the loss function leading to different robust estimators,

$$\text{AIC}_\rho = 2\sum_{i=1}^{n} \rho(Y_i|x_i, \widehat{\theta}) + 2\ \text{Trace}(J_n^{-1}K_n) \tag{9}$$

and select that model which has the smallest AIC_ρ value.

3.2 AIC for M and MM-estimators

Ronchetti (1997) states a robust AIC for M-estimators, which fits within the form (9),

$$\text{AIC}_\rho.\text{M} = 2\sum_{i=1}^{n} \rho_c\left(\frac{Y_i - \widehat{\theta}_m^t x_i}{\widehat{\sigma}_m}\right) + 2\ \text{Trace}(J_{m,n}^{-1}K_{m,n}), \tag{10}$$

where ρ_c is, for example, the Huber loss function as in (4), $\widehat{\omega}_m = (\widehat{\theta}_m, \widehat{\sigma}_m)$, $\widehat{\theta}_m$ and $\widehat{\sigma}_m$ are M estimators, with empirical information matrices $J_{m,n} = -\sum_{i=1}^{n} \frac{\partial\psi(y_i|x_i,\widehat{\theta}_m)}{\partial\widehat{\omega}_m}$ and $K_{m,n} = \sum_{i=1}^{n} \psi(y_i|x_i, \widehat{\theta}_m)\psi^t(y_i|x_i, \widehat{\theta}_m)$, with ψ the derivative of ρ_c with respect to $\widehat{\omega}_m$.

For MM and uniform MM-estimators, the following forms of robust AIC versions are obtained in a similar fashion, where the information matrices can be gotten from the corresponding expressions for S estimators (see Section 3.3) by replacing ρ_0 by $\tilde{\rho}_d$,

$$\text{AIC}_\rho.\text{MM} = 2\sum_{i=1}^{n} \tilde{\rho}_d\left(\frac{Y_i - \widehat{\theta}_{mm}^t x_i}{\widehat{\sigma}_{mm}}\right) + 2\ \text{Trace}(J_{mm,n}^{-1}K_{mm,n}), \tag{11}$$

$$\text{AIC}_\rho.\text{UMM} = 2\sum_{i=1}^{n} \tilde{\rho}_d\left(\frac{Y_i - \widehat{\theta}_{mm}^t x_i}{\widehat{\sigma}_{mm}}\right) + 2\ \text{Trace}(J_{mm,n}^{-1}K_{umm,n}). \tag{12}$$

Again, the smallest such value points towards the preferred model.

3.3 AIC based on robust scale estimators

For S estimators the above approach does not work because of the constraint (5). Therefore, based on (2), we propose a robust AIC with respect to S-estimation of the following form

$$\text{AIC.S} = 2\,n\log(\widehat{\sigma}_s) + 2\ \text{Trace}(J_{s,n}^{-1}K_{s,n}). \tag{13}$$

In this criterion we use the robust S-scale estimator $\widehat{\sigma}_s$ and take possible model misspecification into account by the form of the penalty term (rather than just counting the number of parameters). The empirical information matrices $J_{s,n}$ and $K_{s,n}$ are defined as follows,

$$J_{s,n} = \frac{1}{n}\sum_{i=1}^{n} \rho_d''\left(\frac{y_i - \widehat{\theta}_s^t x_i}{\widehat{\sigma}_s}\right)\frac{x_i x_i^t}{\widehat{\sigma}_s^2} \text{ and } K_{s,n} = \frac{1}{n}\sum_{i=1}^{n} \rho_d'^2\left(\frac{y_i - \widehat{\theta}_s^t x_i}{\widehat{\sigma}_s}\right)\frac{x_i x_i^t}{\widehat{\sigma}_s^2}.$$

Model selection proceeds by computing AIC.S for all models under consideration and by selecting the model with the highest value of AIC.S.

When $\rho(t) = t^2$, this criterion reduces to Takeuchi's information criterion TIC (Takeuchi, 1976) for normal data.

For uniform S estimators, Omelka and Salibián-Barrera (2008) show that, under certain regularity conditions, $\widehat{\sigma}_s$ and $\widehat{\theta}_s$ are asymptotically normally distributed uniformly over the contamination neighbourhood. In this case it is shown that $\sqrt{n}(\widehat{\theta_s} - \theta_s(H)) \sim N_p(0, \Sigma_H)$, with $\Sigma_H = J_{us}^{-1} K_{us} J_{us}^{-1}$ and

$$\begin{aligned}
K_{us} &= E_H[\rho_0'^2(u_1(H))\frac{X_1 X_1^t}{\sigma_s(H)^2}] + \frac{d_H}{b_H}\frac{d_H^t}{b_H}\mathrm{E}_H[(\rho_0(u_1(H)) - b)^2] \\
&\quad -E_H[\rho_0'(u_1(H))(\rho_0(u_1(H)) - b)X_1^t]\frac{d_H^t}{\sigma_s^2(H)} - \frac{d_H}{b_H}\mathrm{E}_H[\rho_0'(u_1(H))(\rho_0(u_1(H)) - b)\frac{X_1^t}{\sigma_s(H)}], \\
J_{us} &= \mathrm{E}_H\left[\rho_0''\left(\frac{Y_1 - \theta_s(H)^t X_1}{\sigma_s(H)}\right)\frac{X_1 X_1^t}{\sigma_s^2(H)}\right],
\end{aligned}$$

where E_H means that the expectation is computed with respect to H and

$$\begin{aligned}
d_H &= E_H\left[\rho_0''\left(\frac{Y_1 - \theta_s(H)^t X_1}{\sigma_s(H)}\right)\frac{(Y_1 - \theta_s(H)^t X_1)X_1^t}{\sigma_s(H)^2}\right] \\
b_H &= E_H\left[\rho_0'\left(\frac{Y_1 - \theta_s(H)^t X_1}{\sigma_s(H)}\right)\frac{(Y_1 - \theta_s(H)^t X_1)}{\sigma_s(H)}\right] \\
u_1(H) &= \frac{Y_1 - \theta_s(H)^t X_1}{\sigma_s(H)}.
\end{aligned}$$

For the calculations of the penalty term in the robust AIC, we use the corresponding empirical information matrices, where $J_{us,n}$ is equal to $J_{s,n}$. Hence, the difference lies in the asymptotic variance component $K_{us,n}$, which results in a larger variance for uniform S-estimators by taking the contamination neighborhoods into account. This leads immediately to a robust AIC based on uniform asymptotic results for S-estimators,

$$\text{AIC.US} = 2\,n\log\widehat{\sigma}_s + 2\ \text{Trace}(J_{s,n}^{-1}K_{us,n}), \tag{14}$$

where, for example, $\rho_0 = \rho_d$ is Tukey's bi-square loss function.

In this same spirit, we propose robust AIC versions based on M-, MM- and UMM-estimators as follows,

$$\text{AIC.M} = 2\,n\log(\widehat{\sigma}_m) + 2\ \text{Trace}(J_{m,n}^{-1}K_{m,n}), \tag{15}$$

$$\text{AIC.MM} = 2\,n\log(\widehat{\sigma}_m m) + 2\ \text{Trace}(J_{mm,n}^{-1}K_{mm,n}), \tag{16}$$

$$\text{AIC.UMM} = 2\,n\log(\widehat{\sigma}_u mm) + 2\ \text{Trace}(J_{umm,n}^{-1}K_{umm,n}), \tag{17}$$

The model with the smallest AIC value points towards the preferred model. Our simulation studies show that these robust scale based-criteria lead to a better performance as compared to the versions (10)–(12).

4 Numerical results

4.1 Simulation settings

The settings for the simulation study are as follows. Design variables $X_1, \ldots, X_p$ are generated from a multivariate normal distribution with mean vector $\mu = (1, \ldots, p)$ and variance covariance matrices (i) a $(p \times p)$ identity matrix for independent Xs and (ii) for dependent Xs, we used Σ_1 and Σ_2 for the cases $p = 6$ and $p = 10$ respectively, where Σ_2 is partitioned in a 6×6 block in the upper left corner, and a 4×4 block in the lower right corner,

$$\Sigma_1 = \begin{pmatrix} 1.0 & 0.6 & 0.6 & 0.05 & 0.05 & 0.05 \\ 0.6 & 1.0 & 0.6 & 0.05 & 0.05 & 0.05 \\ 0.6 & 0.6 & 1.0 & 0.05 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.05 & 1.0 & 0.3 & 0.3 \\ 0.05 & 0.05 & 0.05 & 0.3 & 1.0 & 0.3 \\ 0.05 & 0.05 & 0.05 & 0.3 & 0.3 & 1.0 \end{pmatrix}, \Sigma_2 = \left(\begin{array}{cccc|cccc} 1.0 & 0.6 & \cdots & 0.6 & 0.05 & \cdots & \cdots & 0.05 \\ 0.6 & \ddots & \ddots & \vdots & \vdots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & 0.6 & \vdots & & \ddots & \vdots \\ 0.6 & \cdots & 0.6 & 1.0 & 0.05 & \cdots & \cdots & 0.05 \\ \hline 0.05 & \cdots & \cdots & 0.05 & 1.0 & 0.4 & 0.4 & 0.4 \\ \vdots & \ddots & & \vdots & 0.4 & 1.0 & 0.4 & 0.4 \\ \vdots & & \ddots & \vdots & 0.4 & 0.4 & 1.0 & 0.4 \\ 0.05 & \cdots & \cdots & 0.05 & 0.4 & 0.4 & 0.4 & 1.0 \end{array}\right)$$

These values are kept fixed for all settings to reduce simulation variability. We took sample sizes equal to 50 and 100. Since the results were quite similar, we here only show the results for the sample size equal to 50. We have fitted all $2^p - 1$ possible models without interactions with these p variables.

For the mean structure, we have used the functions $m_1(x) = 1 + x_1 + x_2 + x_3$ for the setting with $p = 6$ and $m_2(x) = 1 + \sum_{j=1}^{6} x_j$ for the setting with $p = 10$, with $x = (x_1, \ldots, x_p)$. As error distribution we used $N(0, 0.7^2)$. We compare nine different AIC versions in this simulation study: classical AIC based on maximum likelihood estimation assuming a normal distribution, the scale based versions (13)–(17), as defined in Section 3.3, and the versions using the ρ-function of (10)–(12) of Section 3.2.

To compute the robust M, S and MM-estimators, we used, respectively, the functions `rlm()`, `lmrob.S()` and `lmrob..M..fit()` from the `R` libraries `MASS` and `robustbase`. In order to investigate the robustness of the methods against outliers, we considered three situations: (i) outliers on the response only, (ii) outliers on the response and some of the covariates, and (iii) outliers on some of the covariates only. For case (i) we randomly generated different percentages of outliers (0%, 5%, 10%, 20%, 30%, 40% and 50%) from $N(50, 0.1^2)$ for each of the simulated cases. For case (ii) we considered the same setting as for (i) for the outlying response variables and in addition generated 30% outliers on the significant variables X_1 and X_2 from a $N(50, 5^2)$ distribution. For case (iii) different percentages of outliers (0%, 5%, 10%, 20%, 30%) on the variables X_1 and X_2 are generated from a $N(500, 5^2)$ distribution. For each of these settings we simulated 1000 samples.

4.2 Simulation results

A summary of the simulation results is provided by reporting the proportions of selected models that are

(O) Overfit - Models containing all the variables in the true model plus some more that are actually redundant.

(C) Correct fit - The true model only.

(U) Underfit - Models with only a strict subset of the variables in true model.

(W) Wrong fit - All models that are not overfit (O), not a correct fit (C) nor underfit (U). These are the models where some of the relevant variables might be present (though not all of them) in addition to some of the redundant variables.

We first consider the case with outlying response values. Table 1 shows detailed simulation results for one of the simulation settings with all AIC methods. As expected, the classical AIC works better than the robust AICs for the data without outliers. Both the classical AIC, $AIC_\rho.M$ and AIC.M select a large proportion of underfit or wrong fit models for the data with outliers, while a higher proportion of overfit and correct fit models are select by AIC.S, AIC.US, AIC.MM and AIC.UMM. All of these methods work better for the cases with a high contamination level of outliers and break down at 50% of outliers in the data; this holds for both dependent and independent Xs. The criteria $AIC_\rho.MM$ and $AIC_\rho.UMM$ select a high proportion of overfit and correct fit models. A comparison of the scale versions to those based on the ρ-function reveals that AIC.MM and AIC.UMM work better than $AIC_\rho.MM$ and $AIC_\rho.UMM$. For the rest of the paper we therefore only present the results using the scale-based versions of the AIC.

Figure 1 shows the results of the proportion of selected correct fit (C) models by different model selection strategies. As expected, the classical AIC works better than the robust AICs for the data without outliers. Both the classical AIC and the AIC.M select only a small proportion of correct fit models, when the data contain outliers on the response variable. That means, both methods are ignoring some of the important variables in the model. A higher proportion of correct fit models is selected by AIC.S for the data set with outliers. This method works better for the cases with a high contamination level of outliers and they break down when there are 50% of outliers in the data.

For small percentages of outliers (10%–20%), the AIC.S method (when tuned to a 50% breakdown point) is not doing well in selecting the correct model. Therefore, we re-compute AIC.S, now tuned to have a 30% breakdown point for the estimators. The corresponding results are plotted in Figure 1 (c) and (d). We observe that this significantly helps for the case of 20% outliers, resulting in a high proportion of correct models selected by AIC.S. When we consider proportion of both overfit and correct fit models together, then AIC.S is performing well for any percentages of outliers with both considered breakdown points. We also computed AIC.US, AIC.MM and AIC.UMM in this simulation setting and observed that the results are similar to those of AIC.S.

The main message to be learned from this simulation study is that AIC based on M estimators using expression (10) is not performing better than just the classical AIC when outliers are present. The AIC versions based on robust scale estimators are preferable. For best performance,

Table 1: Proportion of selected models from classical AIC, AIC with M-estimation (AIC_ρ.M), AIC with MM-estimation (AIC_ρ.MM), AIC with uniform MM-estimation (AIC_ρ.UMM), the robust scale versions: AIC.M, AIC.S, AIC.US, AIC.MM and AIC.UMM. Data are generated with dependent Xs, mean structure m_1 for $p = 6$, error terms from a $N(0, 0.7^2)$ distribution, and for sample size $n = 50$. We consider different percentages ε of outliers generated from $N(50, 0.1^2)$. S- and MM-estimators are computed with 50% breakdown point.

ε		Based on loss function (ρ)				Based on scale estimators				
%		AIC	$AIC_\rho.M$	$AIC_\rho.MM$	$AIC_\rho.UMM$	AIC.M	AIC.S	AIC.US	AIC.MM	AIC.UMM
0	O	0.487	0.091	0.337	0.259	0.249	0.808	0.787	0.821	0.801
	C	0.513	0.020	0.003	0.004	0.085	0.164	0.178	0.159	0.176
	U	0.000	0.179	0.001	0.008	0.115	0.004	0.007	0.005	0.006
	W	0.000	0.710	0.659	0.729	0.551	0.024	0.028	0.015	0.017
5	O	0.002	0.000	0.329	0.255	0.000	0.762	0.755	0.774	0.762
	C	0.001	0.000	0.003	0.003	0.000	0.213	0.217	0.209	0.217
	U	0.552	0.003	0.003	0.006	0.012	0.004	0.008	0.005	0.008
	W	0.445	0.997	0.665	0.736	0.988	0.021	0.020	0.012	0.013
10	O	0.006	0.000	0.293	0.208	0.000	0.738	0.735	0.740	0.738
	C	0.004	0.000	0.003	0.004	0.000	0.235	0.233	0.238	0.238
	U	0.458	0.001	0.002	0.009	0.002	0.007	0.008	0.005	0.006
	W	0.532	0.999	0.702	0.779	0.998	0.020	0.024	0.017	0.018
20	O	0.007	0.006	0.298	0.228	0.001	0.567	0.560	0.565	0.560
	C	0.004	0.000	0.002	0.002	0.000	0.415	0.416	0.418	0.419
	U	0.429	0.002	0.001	0.002	0.008	0.006	0.008	0.006	0.008
	W	0.560	0.992	0.699	0.768	0.991	0.012	0.016	0.011	0.013
30	O	0.012	0.051	0.435	0.370	0.000	0.336	0.340	0.341	0.340
	C	0.005	0.027	0.000	0.000	0.001	0.654	0.651	0.650	0.651
	U	0.413	0.139	0.000	0.000	0.724	0.005	0.005	0.005	0.005
	W	0.570	0.783	0.565	0.630	0.275	0.005	0.004	0.004	0.004
40	O	0.008	0.000	0.828	0.826	0.000	0.075	0.077	0.077	0.077
	C	0.008	0.001	0.000	0.000	0.000	0.906	0.905	0.905	0.905
	U	0.393	0.384	0.000	0.000	0.727	0.014	0.014	0.014	0.014
	W	0.591	0.615	0.172	0.174	0.273	0.005	0.004	0.004	0.004
50	O	0.003	0.000	0.313	0.302	0.000	0.142	0.137	0.138	0.137
	C	0.004	0.000	0.002	0.002	0.002	0.028	0.025	0.027	0.025
	U	0.396	0.504	0.002	0.006	0.396	0.060	0.063	0.061	0.066
	W	0.597	0.496	0.683	0.690	0.602	0.770	0.775	0.774	0.772

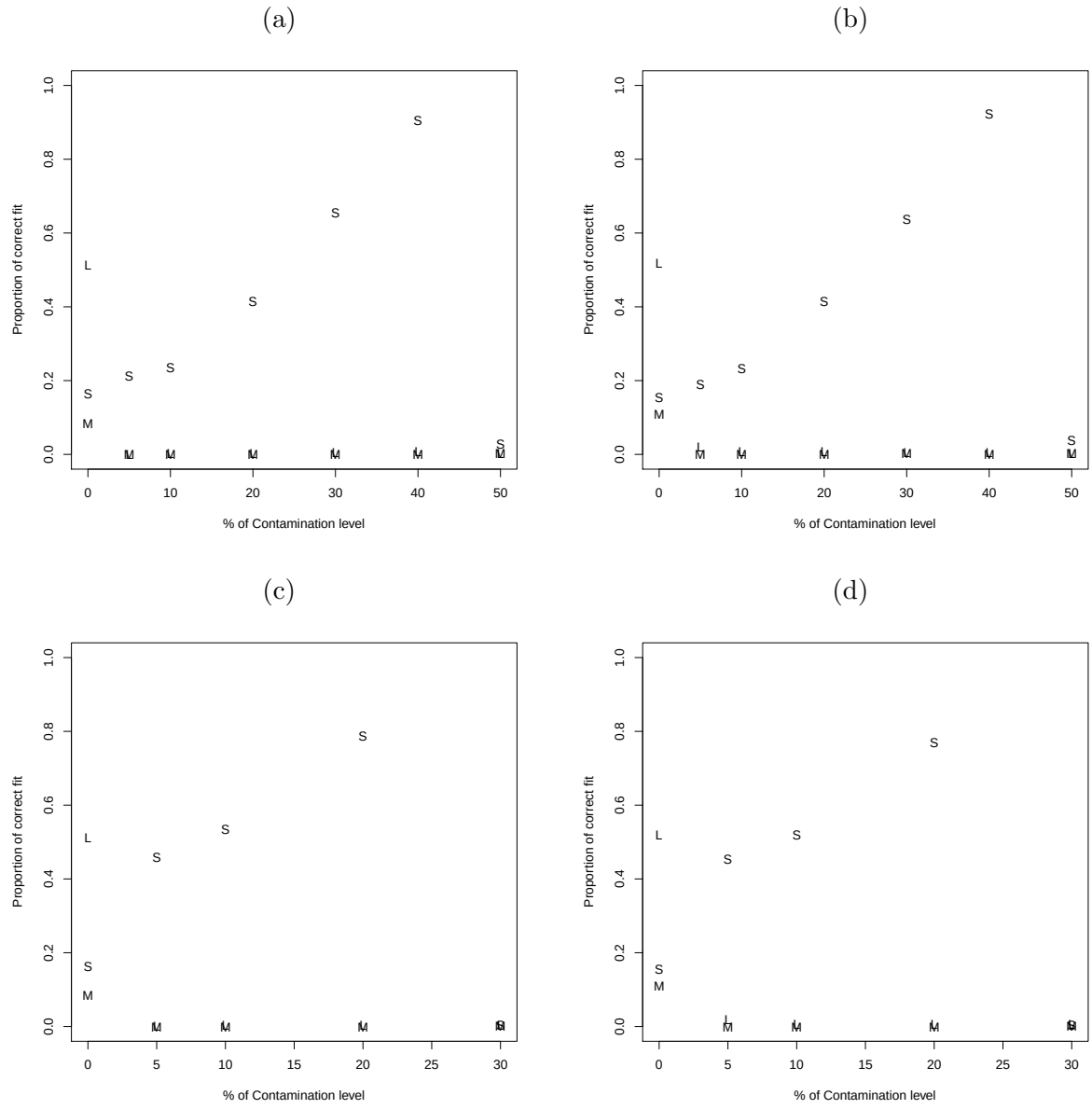

Figure 1: Proportion of selected models from overfit (O) and correct fit (C) from classical AIC (L), AIC based on M-estimators (M) and AIC based on S-estimators (S) for data generated with mean structure m_1 for $p = 6$, error terms from $N(0, 0.7^2)$, sample size $n = 50$ and different percentages of outliers generated from $N(50, 0.1^2)$ for four different cases (a) dependent X with estimators with 50% breakdown point, (b) independent X with estimators with 50% breakdown point, (c) dependent X with estimators with 30% breakdown point, (d) independent X with estimators with 30% breakdown point.

the breakdown point of the estimators should be considered in relation with the proportion of outliers in the data, to avoid underfitting.

More detailed simulation results are shown in Table 2. Again, as expected, the classical AIC works better than the robust AICs for the data without outliers. Both the classical AIC and AIC.M select a large proportion of underfit or wrong fit models for the data with outliers, while a higher proportion of overfit and correct fit models are select by AIC.S, AIC.US, AIC.MM and AIC.UMM. All of these methods work better for the cases with a high contamination level of outliers and break at 50% of outliers in the data; this holds for both dependent and independent Xs. The S estimation based criteria AIC.S and AIC.US give similar results in most of the cases in Table 2. For dependent Xs the proportion of overfit models based on AIC.S and AIC.US is larger than for the case of independent Xs and based on AIC.MM and AIC.UMM is smaller than for the case of independent Xs.

Since table 2 shows that the proportion of selected correct fit models is small for the cases with 5%, 10% and 20% contamination when estimators with 50% breakdown point are used, we recompute the AIC.S, AIC.US, AIC.MM and AIC.UMM for the cases with 0%, 5%, 10%, 20%, 30% contamination level, now with 30% breakdown point estimators. These results are presented in Table 3. It clearly shows that for the case of 20% contamination, the proportion of selected correct fit models is now much larger for the methods AIC.MM and AIC.UMM.

Next, we present results of simulated data with outliers on the explanatory variables, in addition to outliers in the response variable (see the description of case (ii) above). The results are presented in the left part of Table 4. We have fitted all possible models with six explanatory variables in this setting. The results of Table 4 confirm those of Table 3, and illustrate that the investigated robust AIC methods work for both situations.

We also simulated data with different percentage of outliers only on explanatory variables (case (iii)). We compute AIC values from these six different AIC methods and the results are given in right panel of Table 4. Based on these results we observe that the classical AIC method selects a large proportion of overfit and correct fit models for all cases. Therefore, based on this simulation results, it seems valid to use the classical AIC method for the cases with outliers only on the explanatory variables. Also, AIC based on S, uniform S, MM and uniform MM estimators methods select a large proportion of overfit and correct fit models for all cases.

Table 2: Proportion of selected models from classical AIC (AIC), AIC with M-estimation (M), AIC with MM-estimation (MM), AIC with uniform MM-estimation (UMM), AIC with S-estimation (S) and AIC with uniform S-estimation (US) for data generated with independent Xs and dependent Xs, mean structure m_2 for $p = 10$, error terms from a $N(0, 0.7^2)$ distribution, and for sample size $n = 50$. We consider different percentages ε of outliers generated from $N(50, 0.1^2)$. S- and MM- estimators are computed with 50% breakdown point.

ε		Dependent Xs						Independent Xs					
%		AIC	M	S	US	MM	UMM	AIC	M	S	US	MM	UMM
0	O	0.568	0.029	0.600	0.592	0.795	0.780	0.596	0.055	0.399	0.368	0.861	0.851
	C	0.432	0.003	0.055	0.053	0.069	0.077	0.404	0.009	0.029	0.030	0.069	0.074
	U	0.000	0.022	0.006	0.006	0.001	0.001	0.000	0.027	0.016	0.006	0.001	0.001
	W	0.000	0.946	0.339	0.349	0.135	0.142	0.000	0.909	0.556	0.596	0.069	0.074
5	O	0.000	0.000	0.603	0.594	0.769	0.762	0.009	0.000	0.427	0.394	0.839	0.826
	C	0.001	0.000	0.065	0.053	0.089	0.083	0.007	0.000	0.049	0.042	0.086	0.086
	U	0.264	0.274	0.006	0.006	0.000	0.000	0.223	0.170	0.013	0.013	0.000	0.000
	W	0.735	0.726	0.326	0.347	0.142	0.155	0.761	0.830	0.511	0.551	0.075	0.088
10	O	0.001	0.000	0.624	0.606	0.730	0.734	0.000	0.000	0.489	0.463	0.801	0.801
	C	0.000	0.000	0.117	0.120	0.131	0.128	0.001	0.000	0.078	0.079	0.150	0.149
	U	0.316	0.375	0.003	0.004	0.000	0.003	0.264	0.247	0.007	0.003	0.002	0.002
	W	0.683	0.625	0.256	0.270	0.139	0.135	0.735	0.753	0.426	0.455	0.047	0.048
20	O	0.000	0.000	0.628	0.608	0.654	0.652	0.000	0.000	0.591	0.574	0.713	0.714
	C	0.000	0.000	0.227	0.228	0.252	0.252	0.000	0.000	0.211	0.205	0.253	0.252
	U	0.307	0.541	0.000	0.001	0.002	0.002	0.026	0.329	0.005	0.002	0.000	0.000
	W	0.693	0.459	0.145	0.163	0.092	0.094	0.974	0.671	0.193	0.219	0.034	0.034
30	O	0.000	0.000	0.370	0.367	0.365	0.354	0.000	0.000	0.392	0.401	0.389	0.396
	C	0.000	0.000	0.570	0.572	0.579	0.589	0.000	0.001	0.588	0.576	0.604	0.597
	U	0.312	0.456	0.005	0.007	0.003	0.005	0.283	0.455	0.000	0.000	0.000	0.000
	W	0.688	0.544	0.055	0.054	0.053	0.052	0.717	0.544	0.020	0.023	0.007	0.007
40	O	0.000	0.000	0.021	0.021	0.020	0.020	0.000	0.000	0.146	0.145	0.143	0.145
	C	0.000	0.000	0.495	0.496	0.525	0.528	0.000	0.000	0.714	0.710	0.754	0.753
	U	0.314	0.412	0.084	0.083	0.072	0.071	0.289	0.416	0.016	0.016	0.014	0.014
	W	0.686	0.588	0.400	0.400	0.383	0.381	0.711	0.584	0.124	0.129	0.089	0.088
50	O	0.000	0.003	0.023	0.025	0.023	0.023	0.000	0.000	0.031	0.030	0.024	0.026
	C	0.000	0.000	0.002	0.002	0.001	0.001	0.000	0.000	0.000	0.000	0.001	0.001
	U	0.313	0.348	0.017	0.018	0.014	0.016	0.295	0.345	0.015	0.016	0.011	0.010
	W	0.687	0.649	0.958	0.955	0.962	0.960	0.705	0.655	0.954	0.954	0.964	0.963

Table 3: Proportion of selected models from classical AIC (AIC), AIC with M-estimation (M), AIC with MM-estimation (MM) and AIC with uniform MM-estimation (UMM), AIC with S-estimation (S) and AIC with uniform S-estimation (US) for data generated with independent Xs and dependent Xs, mean structure m_2 for $p = 10$, error terms from a $N(0, 0.7^2)$ distribution, and for sample size $n = 50$. We consider different percentages ε of outliers generated from $N(50, 0.1^2)$. S- and MM- estimators are computed with a 30% breakdown point.

ε		Dependent Xs						Independent Xs					
%		AIC	M	S	US	MM	UMM	AIC	M	S	US	MM	UMM
0	O	0.568	0.029	0.601	0.257	0.707	0.710	0.596	0.055	0.398	0.369	0.740	0.730
	C	0.432	0.003	0.057	0.036	0.272	0.267	0.404	0.009	0.028	0.031	0.251	0.261
	U	0.000	0.022	0.007	0.048	0.001	0.001	0.000	0.027	0.017	0.005	0.000	0.000
	W	0.000	0.946	0.335	0.659	0.020	0.022	0.000	0.909	0.557	0.595	0.009	0.009
5	O	0.000	0.000	0.646	0.638	0.655	0.643	0.009	0.000	0.668	0.664	0.679	0.668
	C	0.001	0.000	0.342	0.340	0.338	0.348	0.007	0.000	0.310	0.317	0.318	0.329
	U	0.264	0.274	0.001	0.000	0.000	0.000	0.223	0.170	0.000	0.000	0.000	0.000
	W	0.735	0.726	0.011	0.022	0.007	0.009	0.761	0.830	0.022	0.019	0.003	0.003
10	O	0.001	0.000	0.546	0.533	0.538	0.532	0.000	0.000	0.558	0.558	0.561	0.559
	C	0.000	0.000	0.451	0.461	0.455	0.461	0.001	0.000	0.441	0.442	0.439	0.441
	U	0.316	0.375	0.000	0.000	0.000	0.000	0.264	0.247	0.000	0.000	0.000	0.000
	W	0.683	0.625	0.003	0.006	0.007	0.007	0.735	0.753	0.001	0.000	0.000	0.000
20	O	0.000	0.000	0.177	0.179	0.180	0.179	0.000	0.000	0.196	0.196	0.197	0.196
	C	0.000	0.000	0.820	0.818	0.817	0.818	0.000	0.000	0.804	0.804	0.803	0.804
	U	0.307	0.541	0.001	0.001	0.001	0.001	0.260	0.329	0.000	0.000	0.000	0.000
	W	0.693	0.459	0.002	0.002	0.002	0.002	0.740	0.671	0.000	0.000	0.000	0.000
30	O	0.000	0.000	0.001	0.003	0.000	0.000	0.000	0.000	0.004	0.005	0.002	0.003
	C	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.001	0.000	0.001	0.000
	U	0.312	0.456	0.399	0.380	0.405	0.381	0.283	0.456	0.265	0.181	0.275	0.181
	W	0.688	0.544	0.600	0.617	0.595	0.619	0.717	0.543	0.730	0.814	0.722	0.816

4.3 Employment data in East-central Europe

We used the data set coded ZA3132 from the website http://zacat.gesis.org/webview/index.jsp, named "The evaluation of programs to assist young unemployed in post communist East-Central Europe 1996-1998". We used a subset of this dataset consisting of the response variable, the current monthly earnings (USA $) during 1996-1998, and 16 explanatory variables (see below

Table 4: Proportion of selected models from Classical AIC (AIC), AIC with M-estimation (M), AIC with S-estimation (S), AIC with uniform S-estimation (US), AIC with MM-estimation (MM) and AIC with uniform MM-estimation (UMM) for data generated with dependent X's, mean structure $m_1(x)$ for $p = 6$, error terms from a $N(0, 0.7^2)$, and for sample size $n = 50$. Considered different percentages ε of outliers generated for Y, X_1 and X_2 variables. S- and MM- estimators are computed with a 30% breakdown point.

ε		Different % outliers $\sim N(50, 0.1^2)$ for Y and 30% outliers $\sim N(50, 5^2)$ for X_1 and X_2							No outliers on Y and Different % outliers $\sim N(500, 5^2)$ for X_1 and X_2					
%		AIC	M	S	US	MM	UMM	%	AIC	M	S	US	MM	UMM
0	O	0.491	0.437	0.588	0.577	0.589	0.579	0	0.487	0.249	0.576	0.569	0.578	0.568
	C	0.509	0.138	0.412	0.421	0.411	0.421		0.513	0.085	0.423	0.429	0.421	0.430
	U	0.000	0.113	0.000	0.001	0.000	0.000		0.000	0.115	0.001	0.001	0.001	0.001
	W	0.000	0.312	0.000	0.001	0.000	0.000		0.000	0.551	0.000	0.001	0.000	0.001
5	O	0.076	0.135	0.535	0.524	0.536	0.521	5	0.467	0.416	0.504	0.493	0.786	0.759
	C	0.115	0.017	0.464	0.476	0.464	0.479		0.533	0.131	0.409	0.421	0.206	0.195
	U	0.460	0.210	0.000	0.000	0.000	0.000		0.000	0.106	0.007	0.018	0.002	0.010
	W	0.349	0.638	0.001	0.000	0.000	0.000		0.000	0.347	0.080	0.068	0.006	0.036
10	O	0.022	0.095	0.453	0.450	0.454	0.450	10	0.471	0.205	0.556	0.527	0.589	0.578
	C	0.024	0.015	0.547	0.550	0.546	0.550		0.529	0.077	0.430	0.466	0.411	0.422
	U	0.502	0.417	0.000	0.000	0.000	0.000		0.000	0.134	0.004	0.002	0.000	0.000
	W	0.452	0.473	0.000	0.000	0.000	0.000		0.000	0.584	0.010	0.005	0.000	0.000
20	O	0.004	0.023	0.221	0.219	0.222	0.219	20	0.474	0.037	0.570	0.553	0.595	0.580
	C	0.004	0.004	0.779	0.781	0.778	0.781		0.526	0.007	0.427	0.443	0.405	0.420
	U	0.539	0.239	0.000	0.000	0.000	0.000		0.000	0.089	0.000	0.000	0.000	0.000
	W	0.453	0.734	0.000	0.000	0.000	0.000		0.000	0.867	0.003	0.004	0.000	0.000
30	O	0.007	0.020	0.015	0.018	0.015	0.018	30	0.483	0.078	0.572	0.557	0.568	0.545
	C	0.005	0.023	0.021	0.017	0.021	0.017		0.517	0.011	0.424	0.442	0.432	0.455
	U	0.509	0.375	0.815	0.794	0.811	0.796		0.000	0.177	0.001	0.000	0.000	0.000
	W	0.479	0.582	0.149	0.171	0.153	0.169		0.000	0.734	0.003	0.001	0.000	0.000

for the details). Cases with missing values were removed from the resulting dataset, leading to the subset of 81 observations that we used here.

The explanatory variables are as follows: X_1 age; X_2 gender; X_3 marital status (1-single,2-married/cohabiting, 3-other); X_4 highest level of education (1-less than elementary school, 2-

elementary school, 3-vocational school, 4-professional or technical school, 5-high school/lycee/ gymnasium/grammar school, 6-college, 7-university); X_5 age completed full-time education; X_6 the subject or field specialized in (0-nothing in particular, 1-construction & related, 2-vehicle & machinery repairs, 3-engineering, 4-catering & hospitality, 5-personal & consumer services, 6-health & related, 7-teaching, 8-professional services, 9-other academic subjects); X_7 number of proper jobs since leaving school; X_8 number of holidays away from home during the last 12 months; X_9 amount of time for family; X_{10} amount of time for friends; and X_{11} amount of time for yourself (1-not enough, 2-about right, 3-too much); X_{12} education matches work experience (1-yes(totally), 2-yes(partly), 3-not at all, 4-no work experience); X_{13} use of motor car; X_{14} use of satellite or cable TV; X_{15} use of personal computer; X_{16} use of mobile telephone.

Using standardized residuals plots, it turns out that 17 response values (21%) are outside the range $(-1, 1)$ and can be considered outliers. The use of Cook's distance to the continuous covariates X_1, X_5 and X_8 showed that 9 observations for X_8 can be considered leverage points. We therefore set the S and MM estimation methods to use a 30% breakdown point.

We have fitted all 2^{16} models with a combination of any of these explanatory variables and computed several AIC values for each model. The best three selected models based on each AIC method are given in Table 5.

Table 5: Employment data in East-central Europe. The selected explanatory variables from the classical AIC, AIC with M-estimation, with S-estimation, with uniform S-estimation, with MM-estimation and AIC with uniform MM-estimation, tuned for a 30% breakdown point.

Criteria	Selected variables		
	Best model	Second best model	Third best model
AIC	X_7,X_9,X_{10},X_{12}	X_4,X_5,X_7,X_9	X_7,X_9,X_{12}
AIC.M	X_3	X_3,X_6	X_3, X_{14}
AIC.S , AIC.MM	X_3,X_4,X_7,X_9,X_{10},X_{15}	X_2,X_3,X_4,X_7,X_9,X_{10},X_{15}	X_3,X_4,X_9,X_{10},X_{11},X_{15}
AIC.US , AIC.UMM	X_3,X_4,X_7,X_9,X_{10},X_{15}	X_2,X_3,X_4,X_7,X_9,X_{10},X_{15}	X_2,X_3,X_4,X_9,X_{10},X_{15}

The classical AIC method selects a model with four explanatory variables, while AIC based on M-estimation selects a model with only one variable. For both methods, this number of selected variables is much lower than for the other criteria. This is in line with the simulation results where we observed that AIC and AIC.M have the tendency to select underfit models in

the presence of outliers.

The proposed methods based on S and MM-estimators select the same best model with six variables. Variable X_3, marital status, coincides with the single selected variable from the M-estimation method, The three variables X_7, number of proper jobs, X_9, amount of time for your family and X_{10}, amount of time for friends, are common with the selected set of the non-robust AIC. In addition, the S and MM-based criteria select variables X_4, highest level of education, and X_{15}, use of a PC to explain the current monthly earnings.

4.4 Hofstedt's highway data

We have used Hofstedt's highway data that are available from the R library `alr3` as `data(highway)` (see also Weisberg, 2005). In this dataset there are 39 observations on several highway related measurements. The response variable is the accident rate per million vehicle miles in the year 1973 and there are 11 potential explanatory variables:

X_1 Average daily traffic count(1000's); X_2 Truck volume as a percentage of the total volume;

X_3 Number of lanes of traffic; X_4 Number of access point per mile;

X_5 Number of signalized interchanges/mile; X_6 Number of freeway-type interchanges/mile;

X_7 The speed limit in 1973; X_8 The length of the segments in miles;

X_9 The lane width in feet; X_{10} Width in feet of outer shoulder on the roadway;

X_{11} An indicator of the type of roadway or the source of funding for the road.

We have fitted all 2^{11} possible models with a combination of any of these covariates and computed several AIC values for each model.

We have checked the outliers in this data set using standardized residuals criteria and found that the absolute value of the standardized residuals is larger than 1 for 11 observations. These observations (28% of the data) are considered as outliers. Using Cook's distance, there are 3 observations (cases 25, 34 and 27) with a large value.

The classical AIC selects a model with four explanatory variables, see Table 6, and thus omits seven potential explanatory variables. The robust model selection strategies as given in this paper select models with more variables. AIC based on M-estimation selects a model with five variables. For this example, two of the selected variables coincide with those of AIC, the other three are different. All of the best five models based on AIC and AIC.M contain only few variables (3, 4 or 5 variables based on AIC and 4 or 5 variables based on AIC.M). Table 7

Table 6: Highway data. The selected explanatory variables from the classical AIC, AIC with M-estimation, with S-estimation, with uniform S-estimation, with MM-estimation and AIC with uniform MM-estimation, tuned for a 30% breakdown point.

Criteria	Selected variables
AIC	X_4,X_5,X_7,X_8
AIC with M-estimator	X_2,X_4,X_5,X_{10},X_{11}
AIC with S-estimator	X_2,X_3,X_5,X_6,X_7,X_8,X_{11}
AIC with uniform S-estimator	X_2,X_3,X_5,X_6,X_7,X_8,X_{11}
AIC with MM-estimator	X_3,X_4,X_8,X_9,X_{10},X_{11}
AIC with uniform MM-estimator	X_3,X_4,X_8,X_9,X_{10},X_{11}

presents the five best models as ranked by their AIC values, using AIC with S, uniform S, MM and uniform MM-estimators. The corresponding ranks given by AIC and AIC.M are large for these same models, indicating low preference. AIC.S and AIC.US select the same best model with seven variables, this is for the situation where the breakdown point of the estimators is tuned to 30% to accommodate the about 28% of outliers in the data. The model selected by AIC.MM and AIC.UMM corresponds to the 4th ranked model by AIC.S and contains six variables.

Table 7: Highway data. The selected explanatory variables from highest ranked models based on AIC.S, AIC.US, AIC.MM and AIC.UMM using estimators with 30% breakdown point.

Variables in the selected models											Number of	Rank of	Rank of	Rank of	Rank of
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	variables	AIC.S	AIC.US	AIC.MM	AIC.UMM
0	1	1	0	1	1	1	1	0	0	1	7	1	1	1101	1059
1	1	0	0	1	0	1	0	1	1	1	7	2	2	684	636
0	1	1	1	0	0	0	1	1	0	0	5	3	3	216	195
0	0	1	1	0	0	0	1	1	1	1	6	4	4	1	1
0	0	0	1	1	0	0	1	0	0	1	4	5	5	2	2
0	0	1	1	0	0	0	1	1	1	0	5	6	7	3	4
1	0	1	1	1	0	0	1	0	0	1	6	7	9	4	6
0	0	1	1	1	0	0	1	0	0	1	5	8	6	5	3

5 Discussion

In this paper the definition of the AIC is extended to be used with S and MM estimators.

It turns out that the classical (non-robust) AIC works well for data sets with only few outlying observations, and with data where the outliers are only in the explanatory variables. The use of AIC based on M-estimation is not encouraged, based on our simulation results, since with this method important variables might be ignored in the selected model, resulting in underfit models. The versions of AIC that use robust scale estimators arising from S and MM-estimators come out best in the comparison. For these methods, the breakdown point of the estimation method should be tuned in accordance with the percentage of outliers in the data. These methods are particularly useful when there are outliers on the response variable.

In line with the known properties of the non-robust AIC, these versions of AIC based on S and MM-estimators, have the tendency to slightly overfit, which ensures that no important variables are lost when this method is used as a screening step to indicate potential important variables in a full analysis of the data. In our simulation studies, the average number of redundant variables in overfitted models was between 2 and 3.

While our study has focussed on the AIC as a variable selection tool, it might be of interest to extend other robust variable selection methods that currently mainly deal with M-estimators, to more advanced robust estimation methods, such as S or MM-estimators.

Acknowledgements

We thank C. Agostinelli for helpful comments on an earlier version of this paper.

A Appendix: R-code

Define AIC function - S and MM estimator

```
AIC.S<- function(y, X, beta.s,scale.s, cc=1.54764)
{
U=matrix(ncol=1,nrow=n)
UU=matrix(ncol=1,nrow=n)
UUU=matrix(ncol=1,nrow=n)
for(i in 1:n){
U[i,]=dPsi((y[i,]-X[i,] %*% beta.s)/scale.s ,cc)
```

```
UU[i,]=(Psi((y[i,]-X[i,] %*% beta.s)/scale.s ,cc))^2
}
J= (t(X) %*% diag(as.vector(U)) %*% X * (1/(scale.s^2)))/n
inv.J<- solve(J)
K= (t(X) %*% diag(as.vector(UU)) %*% X * (1/(scale.s^2)))/n
AIC =2*n*(log(scale.s))+ 2* sum(diag(inv.J %*%(K)))
return(AIC)
}

Rho = function(x, cc){
    U = x/cc;  U1 = 3 * U^2 - 3 * U^4 + U^6
    U1[abs(U) > 1] = 1;   return(U1)}

Psi = function(x, cc){
   U = x/cc;  U1 = 6/cc * U * (1 - U^2)^2
    U1[abs(U) > 1] = 0;   return(U1)}

dPsi = function(x, cc){
    U = x/cc;  U1 = (6/(cc^2)) *(1- 6* U^2+ 5* U^4)
    U1[abs(U) > 1] = 0;   return(U1)}
```

Define AIC function - uniform S and MM estimator

```
AIC.US<- function(y, X, beta.s,scale.s, cc=1.54764, b=0.5)
{
U=matrix(ncol=1,nrow=n)
UU=matrix(ncol=1,nrow=n)
UU2=matrix(ncol=1,nrow=n)
UUUb=matrix(ncol=1,nrow=n)
UUUb2=matrix(ncol=1,nrow=n)
UUUU=matrix(ncol=1,nrow=n)
BH1=matrix(ncol=1,nrow=n)
DH1=matrix(ncol=1,nrow=n)
for(i in 1:n){
U[i,]=dPsi((y[i,]-X[i,] %*% beta.s)/scale.s ,cc)
UU[i,]=Psi((y[i,]-X[i,] %*% beta.s)/scale.s ,cc)
UU2[i,]=(Psi((y[i,]-X[i,] %*% beta.s)/scale.s ,cc))^2
UUUb[i,]=Rho(((y[i,]-X[i,] %*% beta.s)/scale.s) , cc) - b
UUUb2[i,]=(Rho(((y[i,]-X[i,] %*% beta.s)/scale.s) , cc) - b)^2
UUUU[i,] = UU[i,]*UUUb[i,]
```

```
BH1[i,]=(UU[i,] * ((y[i,]-X[i,] %*% beta.s)/scale.s))
DH1[i,]=U[i,] * ((y[i,]-X[i,] %*% beta.s)/scale.s)
}
Vsi= (t(X)%*%X)/(scale.s^2)
Jsi= (t(X) %*% diag(as.vector(U)) %*% X * (1/(scale.s^2)))/n
inv.Jsi<- solve(Jsi)
dh=sum(DH1)/n * as.matrix(X/scale.s)
bh=sum(BH1)/n
dbh=dh/bh
E1=(t(X) %*% diag(as.vector(UU2)) %*% X * (1/(scale.s^2)))/n
E2=(t(dbh) %*% diag(as.vector(UUUb2)) %*% dbh * (1/(scale.s^2)))/n
E3=((t(X)/scale.s) %*% diag(as.vector(UUUU)) %*% dbh)/n
E4=(t(dbh) %*% diag(as.vector(UUUU)) %*% (X/scale.s))/n
Ksi= E1+E2-E3-E3
AIC =2*n*(log(scale.s)) + 2* sum(diag(inv.Jsi %*%Ksi))
return(AIC)
}
```

Define AIC function - M estimator

```
AIC.M<- function(y, X, beta.m,scale.m, cval=1.345)
{
U=matrix(ncol=1,nrow=n)
UU=matrix(ncol=1,nrow=n)
UUU=matrix(ncol=1,nrow=n)
SC1=matrix(ncol=1,nrow=n)
SC2=matrix(ncol=1,nrow=n)
for(i in 1:n){
U[i,]=dPsiM((y[i,]-X[i,] %*% beta.m)/scale.m ,cval)
UU[i,]=(PsiM((y[i,]-X[i,] %*% beta.m)/scale.m ,cval))^2
UUU[i,]=-RhoM((y[i,]-X[i,] %*% beta.m)/scale.m ,cval)
SC1[i,]= (-(UU[i,] * ((y[i,]-X[i,] %*% beta.m)/scale.m^2)))^2
SC2[i,]=U[i,]*((y[i,]-X[i,]%*%beta.m)^2/scale.m^4)+(UU[i,]*((y[i,]-X[i,]%*% beta.m)
/scale.m^3)) }
SC.c=matrix(0,ncol(X),1)
SC.r=matrix(0,1,ncol(X))
J.beta=( t(X) %*% diag(as.vector(U)) %*% X * (1/(scale.m^2)))/n
J=rbind(cbind(J.beta,SC.c),cbind(SC.r,(sum(SC2)/n)))
inv.J<- solve(J)
K.beta= (t(X) %*% diag(as.vector(UU)) %*% X * (1/(scale.m^2)))/n
```

```
K=rbind(cbind(K.beta,SC.c),cbind(SC.r,(sum(SC1)/n)))
AIC =2*n*log(scale.m)+ 2 *sum(diag(inv.J %*%K))
#return(list(AIC,sum(diag(inv.J %*%K))))
return(AIC)
}

RhoM = function(x, cval){
     Rho1 = ifelse(abs(x)<=cval,(x^2),(2*cval*abs(x)-cval^2)); return(Rho1)}
PsiM = function(x, cval){
     Psi1 = ifelse(abs(x)<=cval,2*x,2*cval*sign(x)); return(Psi1)}
dPsiM = function(x, cval){
     dPsi1 = ifelse(abs(x)<=cval,2,0); return(dPsi1)}
```

References

Agostinelli, C. (2002). Robust model selection in regression via weighted likelihood methodology. *Statistics and Probability Letters*, 56:289–300.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. and Csáki, F., editors, *Second International Symposium on Information Theory*, pages 267–281. Akadémiai Kiadó, Budapest.

Beaton, A. and Tukey, J. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16:147–185.

Claeskens, G. and Hjort, N. (2008). *Model Selection and Model Averaging*. Cambridge University Press, Cambridge, UK.

Heritier, S., Cantoni, E., Copt, S., and Victoria-Feser, M.-P. (2009). *Robust methods in biostatistics*. Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chichester.

Huber, P. (2004). *Robust Statistics*. John Wiley and Sons.

Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika*, 83(4):875–890.

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust Statistics Theory and Methods*. John Wiley and Sons.

Müller, S. and Welsh, A. (2005). Outlier robust model selection in linear regression. *Journal of the American Statistical Association*, 100:1297–1310.

Omelka, M. and Salibián-Barrera, M. (2008). Uniform asymptotics for S- and MM-regression estimators. *The Annals of the Institute of Statistical Mathematics*. To appear.

Qian, G. and Künsch, H. R. (1998). On model selection via stochastic complexity in robust linear regression. *Journal of Statistical Planning and Inference*, 75:91–116.

Ronchetti, E. (1985). Robust model selection in regression. *Statistics and Probability Letters*, 3:21–23.

Ronchetti, E. (1997). Robustness aspects of model choice. *Statistica Sinica*, 7:327–338.

Ronchetti, E., Field, C., and Blanchard, W. (1997). Robust linear model selection by cross-validation. *Journal of the American Statistical Association*, 92:1017–1023.

Ronchetti, E. and Staudte, R. G. (1994). A robust version of Mallows' C_p. *Journal of the American Statistical Association*, 89:550–559.

Rousseeuw, P. and Yohai, V. (1984). Robust regression by means of S-estimators. In *Robust and nonlinear time series analysis (Heidelberg, 1983)*, volume 26 of *Lecture Notes in Statist.*, pages 256–272. Springer, New York.

Salibián-Barrera, M. and Van Aelst, S. (2008). Robust model selection using fast and robust bootstrap. *Computational Statistics and Data Analysis*, 52:5121–5135.

Sommer, S. and Staudte, R. G. (1995). Robust variable selection in regression in the presence of outliers and leverage points. *Australian Journal of Statistics*, 37:323–336.

Takeuchi, K. (1976). Distribution of informational statistics and a criterion of model fitting. *Suri-Kagaku (Mathematical Sciences)*, 153:12–18. In Japanese.

Weisberg, S. (2005). *Applied Linear Regression*. John Wiley and Sons Inc., Hoboken, New Jersey., 3rd edition.